

From corpus to grammar: how DOBES corpora can be exploited for descriptive linguistics

Peter Bouda[♦] and Johannes Helmbrecht[♡]

[♦]Ludwig-Maximilians-Universität München, [♡]Universität Regensburg

The principles and techniques of language documentation developed during the last one and half decades and the sheer amount of corpora which have been compiled for endangered languages up to now will have an impact on grammar writing in particular with respect to the data base of grammars. On the other hand, advances in computer technology allow a closer link between corpus data which are the basis for generalizations and the grammatical description itself. The future the grammatical description of a language will not only present selected illustrative examples, but will also be linked to the entire set of corpus data that are the empirical basis for it. This makes generalizations transparent to the reader and open to falsification by the scientific community.

The article critically examines the relations between the DOBES corpus, the analysis and the grammatical description itself. Special attention will be laid on the particular the two fundamental perspectives of a semasiological and an onomasiological grammar, can be translated into the various kinds of search and concordancing routines to be executed in the corpus analysis. We present a typology of searches descriptive linguists need to apply. This typology defines requirements with regard to the functionality of specific software to be developed. In the second part, the article presents a technical solution, a preliminary version of a database/concordancing software specifically designed to fulfill the functions and principles outlined in the preceding sections.

1 INTRODUCTION Right now, there are about 50 DOBES projects documenting endangered languages around the world (cf. the DOBES map below).

Language documentation is a multi-purpose enterprise, but one important purpose of DOBES corpora is to serve as data bases for grammatical descriptions of these languages. Since these corpora have a different structure and different properties than the large monolingual corpora that are available for English, German and other major European languages they allow for and require different corpus linguistic methods. Or, to put it the other way round, corpus linguistic methods have to be adapted to the specific properties of DOBES corpora.

One of the great advantages of DOBES corpora over these traditional corpora is that they are always bilingual with an idiomatic translation in one of the main European languages, and in addition, that they often include a morpheme-by-morpheme glossing. This means



FIGURE 1.: The DOBES documentation projects <http://www.mpi.nl/dobes>

that DOBES corpora contain much more additional information in form of these types of annotations than traditional corpora. The large monolingual corpora mostly consist of plain text only. It is the overall goal of this article to show how corpora such as DOBES corpora can be exploited by corpus linguistic means in order to extract the kind of data that are necessary for grammatical descriptions.

What kind of data is necessary for a grammatical description can be deduced from the principles and requirements of grammaticography. One important distinction that is maintained in grammaticography is the one between a semasiological and an onomasiological approach to grammar.

The semasiological approach takes the formal expressions of a language as starting point and tries to find out what they mean in different contexts. This approach resembles the analytic tasks hearers have to fulfill in interpreting speakers' utterances. They have to start from the uttered expression and have to assign a meaning to them which is appropriate in the given shared context of the utterance.

The onomasiological approach, on the other hand, takes the meaning/ function as starting point and tries to find out which forms in a language may be used to express them. This approach resembles the generative task the speaker has to fulfill in order to form an utterance. The speaker has a concept of what he/she wants to communicate to the hearer and his/her task is to find the appropriate expression in a given shared context that fulfills the intended meaning best.

These different but complementary approaches to grammar have also an impact on the data searching methods to be applied to the corpus. The semasiological approach dominates in traditional computational treatments of monolingual text corpora. Specific lexical or grammatical forms in the target language are searched for in digital corpora. The goal is to

collect all different contexts in which these forms appear in order to find out the conventional meanings of these forms and their variation in different contexts.

The onomasiological approach, on the other hand, requires a semantic annotation such that all forms and constructions that express a certain semantic notion – for instance possession – are annotated in a way that the search for the annotation “possession” gives all different expressions of this notion. A semantic annotation is – of course – the rare exception in digital corpora, since they have to be added manually, which is a quite time-consuming process.

It is one of the main goals of this paper to show that the particular strength of DOBES corpora is that they can be exploited much more systematically than traditional monolingual corpora. They allow for searches that provide the necessary data for both the semasiological and the onomasiological approach in direct and indirect ways. The second goal of this paper is to present a linguistic database and concordancing software – the Poio Analyzer – that allows easily – i.e. in a user friendly manner – to conduct the text searches that are necessary to extract the kinds of data from a DOBES corpus that are required for the two different analytical approaches to grammar just mentioned.

The structure of the paper is as follows. The main properties of DOBES corpora and their implications for the various search types will be dealt with in Section 2. Section 3 presents a general typology of possible search types within DOBES corpora. These search types are – in principle – independent of the two approaches to grammar and the types of data they require. It will be demonstrated that these kinds of searches exceed the possibilities that monolingual corpora allow for. In Section 4, it will be exemplified how these search types can be utilized to gain data that are relevant for a semasiological description. In Section 5 in turn, it will be exemplified how data may be extracted from a DOBES corpus – using these search types – that are relevant for the onomasiological description. In Section 6 and Section 8, the functionality and the shortcomings of Elan with regard to concordances will be evaluated and the prototype of a linguistic database and concordancing software – the Poio Analyzer – will be introduced that is designed specifically for the needs of a descriptive linguist.

2 THE STRUCTURE AND PROPERTIES OF DOBES CORPORA The typical structure of an annotated text of a DOBES corpus can be seen in Figure 2. An annotated text consists minimally: a) of a text tier (abbreviated *tx* in Figure 2) containing the transcribed text of the audio and video recording, and b) a free idiomatic translation tier (abbreviated *ft* in Figure 2). Many DOBES corpora have, in addition, a morpheme-by-morpheme glossing such that there is c) a tier with a morpheme segmentation (abbreviated *mo* in Figure 2) and d) a tier with lexical and grammatical glosses corresponding to the segmented morphemes in the *mo* tier (abbreviated *gl* in Figure 2).

The *tx* tier contains the usually time-aligned transcription of the audio/video recording. These transcriptions are usually not based on the IPA, but on an already existing conventional orthography, or on a practical orthography developed by the documentation team. Associated with the *tx* tier there is a *ft* tier which contains the free or idiomatic translation of the target language text. The language used for the free translation is usually one of the major European languages often the one that replaces the endangered language in the community. In the example in Figure 2, this is English. Many documentation teams enriched the

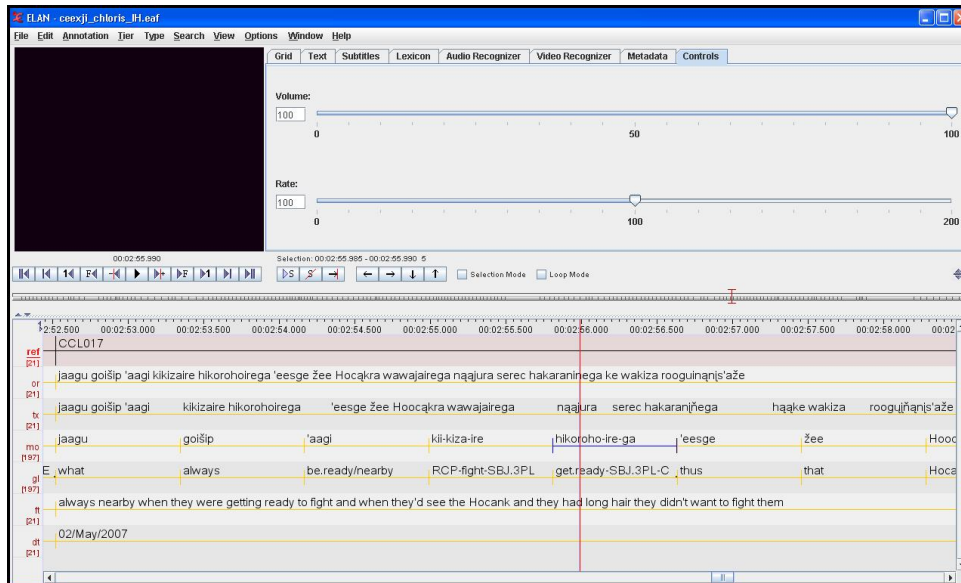


FIGURE 2.: Screen shot of an annotated Hocak (Siouan) sentence in Elan

transcription in the tx tier and its translation in the ft tier with a second type of translation called lexical and grammatical morpheme-by-morpheme glossing. For this, a morpheme segmentation in the mo tier is matched with a gl tier that contains the lexical or grammatical glosses; cf. Figure 2.

It has to be kept in mind that these four essential parts of an annotated text involve previous and preliminary decisions on the side of the linguist that may turn out to be wrong later on in the process of the advancing documentation.

(1) With regard to the text tier (tx):

1. The practical orthography developed by the documentation team may be based on a wrong phonological analysis;
2. The segmentation of the text in sentence/ utterance units may turn out to be wrong later on;
3. The segmentation of the text into words may turn out to be wrong later on;

(2) With regard to the translation tier (ft):

1. The free translation into the European language may be semantically incomplete (i.e. certain forms of the target language are not translated and are left to inference) or may contain semantic elements that are only implied in the text of the target language but are obligatory in the mediating language;

(3) With regard to the morpheme and gloss tiers (mo/gl):

1. The identification of the grammatical and lexical morphemes and their status may turn out to be wrong later on;
2. Or, the assignment of grammatical function/ lexical meaning to a lexical or grammatical form may be wrong or incomplete;

The main goal of the documentation of an endangered language is not the grammatical analysis of the language *per se*, but the collection and processing of a representative corpus of texts in a way that a non-speaker can access it. It is hence obvious that the list of grammatical forms and lexical items has to be preliminary in the documentation process itself. However, even if there are mistakes in the transcriptions, in the translations, and in the morpheme-by-morpheme glossing, all three elements of the annotation represent an invaluable source of additional information that can be exploited in corpus linguistic searches. DOBES corpora are parallel corpora relating a morpheme-by-morpheme translation as well as a free translation to the text of the target language. This kind of additional information presupposes a preliminary phonological analysis (with regard to the transcription) as well as a preliminary grammatical analysis (with regard to the glossing). It will be shown in a moment that digital corpora of this DOBES type allow types of searches that are interesting for descriptive purposes and that go beyond the possibilities of traditional monolingual corpora. These additional search possibilities will be systematically explored in the following section.

3 A TYPOLOGY OF CORPUS LINGUISTIC SEARCHES IN DOBES CORPORA From a technical point of view the search types a descriptive linguist would like to apply to a DOBES corpus are independent of the distinction between the two basic approaches to grammar mentioned above. The most important search types are summarized in Search for a single lexical or grammatical form F₁ or a single construction C₁ consisting of more than one element in a single tier x₁ within a linguistic context defined by the level of syntagmatic complexity/ grammatical level: and Search for the co-occurrence of two or more lexical or grammatical forms F₁ and F₂, or of construction C₁ and C₂ in one specific tier x₁ or across different tiers x₁-n within a linguistic context defined by the level of syntagmatic complexity/ grammatical level:.

- (4) • Search for a single lexical or grammatical form F₁ or a single construction C₁ consisting of more than one element in a single tier x₁ within a linguistic context defined by the level of syntagmatic complexity/ grammatical level:

- a) tier x₁ [...F₁...]word [...C₁-C₁...]word
- b) tier x₁ [...F₁...]phrase [...C₁-C₁...]phrase
- c) tier x₁ [...F₁...]clause/sentence [...C₁-C₁...]clause/sentence

Comments:

- Forms or constructions may be searched within all kinds of tiers, i.e. tier x₁ = {tx, mo, gl, ft, ...}

- F_1 stands for a string of characters that does not exceed the word boundary, e.g. a submorpheme, a morpheme, a word, a lexical or grammatical gloss, a word of the language of translation, etc.}
- C_1 stands for any discontinuous chain of linguistic units with a conventionalized meaning within a certain level of syntagmatic complexity, for instance circumfixes on the word level, or analytic constructions such as the *will* + Verb_{INF} future construction in English;

The Search results should be presented in an interlinear version. It should be possible to store the search results/ hits as a separate data set which may be the basis for another search. It should be possible that the context size of the concordances can be determined with respect to the different levels of syntagmatic complexity (see below Table 1).

Search for the co-occurrence of two or more lexical or grammatical forms F_1 and F_2 , or of construction C_1 and C_2 in one specific tier x_1 or across different tiers x_{1-n} within a linguistic context defined by the level of syntagmatic complexity/ grammatical level:

- a) tier x_1 [... F_1/C_1 - C_1 ... F_2/C_2 - C_2 ...] word/ phrase/ clause/ sentence
Between the linguistic items (i.e. forms/ constructions) F_{1-2}/C_{1-2} which are searched for in this search type – on a single tier x_1 – a logical operation can be defined such F_1 {AND, OR, AND NOT etc.} F_2 , or F_1 {AND, OR, AND NOT etc.} C_1 - C_1 , and so forth.
- b) tier x_1 [... F_1/C_1 - C_1 ...] word/ phrase/ clause/ sentence and tier x_2 [... F_2/C_2 - C_2 ...] word/ phrase/ clause/ sentence
Between the linguistic items F_1 in tier x_1 and F_2 in tier x_2 or C_1 in tier x_1 and C_2 in tier x_2 , and so forth, which are searched for in this search type, a logical operation can be defined such that F_1 in tier x_1 {AND, OR, AND NOT etc.} F_2 in tier x_2 , or C_1 - C_1 in tier x_1 {AND, OR, AND NOT etc.} C_2 - C_2 in tier x_2 , and so forth.

The principle types of corpus linguistic searches summarized in an abstract form in Search for a single lexical or grammatical form F_1 or a single construction C_1 consisting of more than one element in a single tier x_1 within a linguistic context defined by the level of syntagmatic complexity/ grammatical level:-Search for the co-occurrence of two or more lexical or grammatical forms F_1 and F_2 , or of construction C_1 and C_2 in one specific tier x_1 or across different tiers x_{1-n} within a linguistic context defined by the level of syntagmatic complexity/ grammatical level: will be applied to specific data needs of the two complementary approaches to grammar in the subsequent sections (cf. Sections 4 and 5). Different questions in a semasiological approach and an onomasiological approach will be posed and it will be show how these different questions could be translated into different kinds of searches. The illustrating examples come from the Hocak corpus (Hartmann et al. 2009ff), a DOBES corpus compiled within the DOBES project “Documentation of the Hocak Language” some years ago.¹

¹ Cf. the website of the DOBES project “Documentation of the Hocak Language” http://www2.uni-erfurt.de/sprachwissenschaft/Vgl_SW/Hocank/index_frames.html

grammatical units	levels of syntagmatic complexity
submorpheme	word
lexical and grammatical morphemes (stems and affixes) and their combinatory possibilities	word
lexical and grammatical words and their combinatory possibilities	phrase
phrase	clause
clause	(complex) sentence
sentence	text/ discourse

TABLE 1.: Grammatical units and the levels of complexity

4 STRUCTURAL GRAMMAR – THE SEMASIOLOGICAL APPROACH The goal of a structural grammar is to identify the lexical and grammatical units (including grammatical constructions), and to find out the syntagmatic possibilities of the combinations of these units on all syntagmatic levels of complexity. There are at least four such levels of complexity: the word, the phrase, the clause (or simple sentence), and the complex sentence, cf. Table 1.

On the word level, the descriptive linguist has to identify all lexical and grammatical morphemes of a language, their respective paradigms, and all rules of syntagmatic combinations of bound morphemes with lexical morphemes. This task implies among other things that word forms – the primary empirical basis of this task – have to be compared in order to find stem and affixes. This task is easier to solve if there is a preliminary analysis available in form of a morpheme-by-morpheme glossing.

On the phrase level, the descriptive linguist has to identify all lexical and grammatical words and their possible combinations in phrases. This task is easier to solve if there is a parts-of-speech annotation which, unfortunately, is lacking systematically in DOBES corpora.

And last but not least, on the level of the clause and complex sentence, the descriptive linguist has to determine all phrase types and their possible combinations in clauses and complex sentences. This task would be easier to solve, if there were a phrase structure annotation which is lacking in DOBES corpora as well.

One may say that the better or richer the annotations the easier are the problems of a structural grammar to solve. Whether a text corpus comprises a morpheme-by-morpheme glossing or not makes a big difference with regard to the formulation of searches in the text corpus and the exactness and relevance of the hits one gets out of these searches. The subsequent sections illustrate some corpus linguistic methods that can be applied to DOBES corpora in order to gain data that are relevant for a structural grammar and the semasiological approach to grammar. These illustrations remain on the word level. On the higher levels of syntagmatic complexity the lack of a parts of speech and phrase structure annotation in DOBES corpora complicates the application of corpus linguistic methods. The descriptive linguist has to find indirect searches in order to identify the syntactic units a language has above the word level.

xunıg 2	xunınaakšana 1	xunııgra 1
xunııga 1	xunııııge 1	xunııık 3
xunıııgnı 1	xunııık 1	xunıııkra 1
xunııııgnııgre 1	xunııııı 1	xunııııı 5
xunıııııgnııgre 2	xunıııııı 1	xunııııııka 1
xunııııııgra 1	xunııııııık 1	Xunıııııııka 1
xunııııııšana 1	Xurucı 2	xunıııııııııge 1
xunıııık 1	xusgajana 1	xunıııııııııkra 1
xunıııık 1	xuu 3	xunııııııııııı 1
xunıııık 3	xunıııı 3	xunıııııııııııııge 1
Xunııııka 1	xunıııııg 1	xunıııııııııııııııkra 1
xunıııııkra 2	xunııııııııııı 1	xunııııııııııııııııı 1
xunıııııııregi 1	xunııııııııııııka 2	
xunııııııııııı 1	xunıııııııııııııka 1	

TABLE 2.: Extract from the word list of the entire Hocak corpus: *xunıı* ‘small, little’

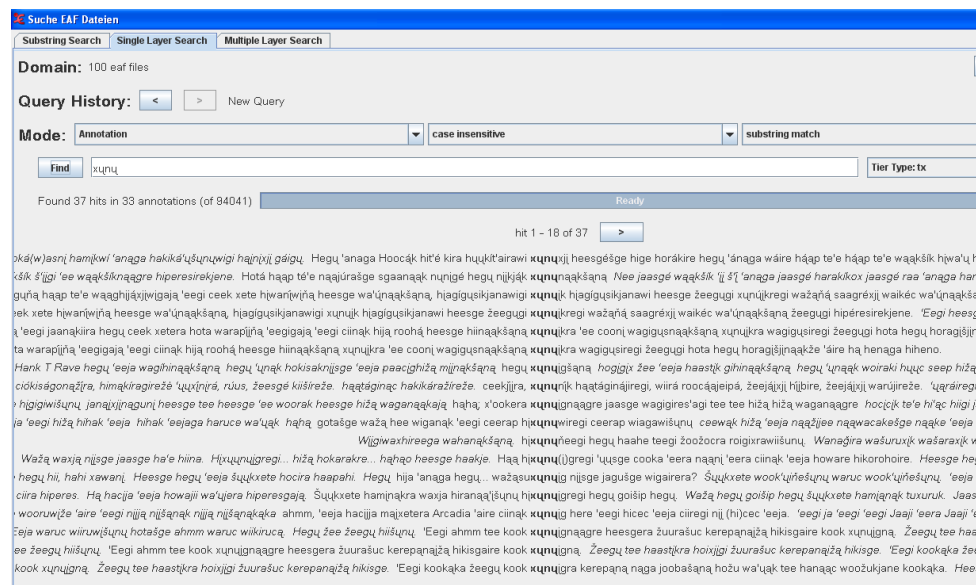
4.1 IDENTIFICATION OF LEXICAL UNITS Words - word forms to be more precise - can be identified in the transcription line (tx tier) by blanks at the left and the right edge. However, problematic cases arise for instance with clitics and compounds. The transcription team necessarily decided at some point in the documentation that a certain form is a clitic, or not, by using blanks. The problem is that this decision may be wrong or inconsistent throughout the corpus for various reasons (e.g. progress or changes in grammatical analysis during the transcription process, inconsistent analysis on the basis of different criteria, native speakers changing intuition etc.). Furthermore, the decision whether a combination of nouns is a coordination of phrases or a nominal compound may be difficult to draw, if there are no morphological markers that indicate this kind of relation; therefore the transcription may be inconsistent with regard to this question too.

The identification of lexical units requires that all forms of a lexeme (including stem allomorphs) and the lexeme itself are identified together with its conventional meaning (polysemy included). Homophonous forms should be found also. There are two possibilities with regard to a DOBES corpus, either the corpus has a morpheme-by-morpheme segmentation/glossing (cf. next subsection Section 4.1.1), or not (cf. Section 4.1.2).

4.1.1 IDENTIFICATION OF LEXICAL UNITS: WITHOUT MORPHEME SEGMENTATION AND GLOSSING IN THE CORPUS There are three possible strategies in order to identify lexical stems:

a) One may export all words of the target language (defined by blanks in the text corpus) and list them alphabetically. Words of the same form are simply counted; the stems of inflected or derived word forms can be found by comparison: the larger stem parts of the words are identical or at least similar. See Table 2 for an example of this procedure from the Hocak corpus.

Table 2 contains a segment of all word forms of the Hocak corpus alphabetically sorted. This wordlist was compiled and exported with Elan. The segment in Table 2 contains all word forms that resemble the lexeme *xunıı* ‘small, little’ in the corpus. As can be seen in this list, there are derived forms of *xunıı* ‘small, little’ with different suffixes, and there are



c) The third strategy starts from the translation tier. One may search for a lexical item in the free translation ft tier assuming that there is a lexical match in the object language. For instance, if you search for *small/little* in the ft tier, you indeed get lexical matches in the target language, cf. (5), or you get forms with a diminutive marker (glossed DIM) as can be seen in (6).

(5) ref ED3024

tx *cii xüünüik hiža 'üü*
 mo *cii xüünü-ik hiža 'üü*
 gl house be.small(OBJ.3SG)-DIM one do/make(SBJ.3SG)

tx *jaagu nji hieci 'eeja*
 mo *jaagu nji hieci*
 gl what water near
 ft he made a small house by the river
 dt 25/Oct/2004

(6) ref CTD003

tx *žee ceek honüwqk 'eeja tee hocicigiža, nijkjak naagre*
 mo *žee ceek ho-nüwqk 'eeja tee hocic-iğ-iža nijkjak naagre*
 gl that first/new APPL.INESS-run there this **boy-DIM-one** child these
 ft from the start, **a little boy**, these children...ahm...
 dt 11/May/2006

In (5), the result of the search is the expected property word *xüünü-ik* 'small-DIM', however, with the diminutive marker *-(n)ik*; in (6), there is no lexical equivalent of the search item 'small', instead one gets a noun with a diminutive marker only. This strategy has the advantage to provide an overview of the forms in the target language that correspond a specific form of the language of translation. It is hence a search type that rather belongs to the onomasiological approach. In addition, it allows the identification of the allomorphs of a morpheme. Another advantage is that this strategy allows the finding of lexical paradigms and stem suppletivism. The disadvantage is that this strategy ignores the polysemy (manifest as variation in the translation) of a form.

4.1.2 WITH MORPHEME-BY-MORPHEME GLOSSING IN THE CORPUS The identification of lexical stems is of course much easier, if there are morpheme breaks and a glossing of the lexical and grammatical morphemes. The morpheme-by-morpheme interlinear glossing presupposes an analysis of lexical and grammatical forms. The mo tier contains the underlying or standardized form of the grammatical and lexical morpheme, and the gl tier attributes the standard glossing (*Grundbedeutung*) of the morpheme. Searches of the lexical item on the mo tier allow finding the allomorphs of a lexical morpheme by comparing the hits with the corresponding form on the tx tier. For instance, in the Hocak corpus, there are quite a few variously abbreviated forms of *anaga* 'and' such as *-naga* in the tx tier which could not be found otherwise. These forms are allomorphs of *anaga*.

4.2 IDENTIFICATION OF GRAMMATICAL UNITS The Identification of grammatical units requires the establishment of the morpheme-allomorph pairings and its grammatical meanings or functions. With respect to corpus linguistic searches in DOBES corpora, there are two possibilities: either the corpus has a morpheme-by-morpheme segmentation/glossing, or not.

4.2.1 IDENTIFICATION OF GRAMMATICAL UNITS: WITHOUT MORPHEME SEGMENTATION AND GLOSSING IN THE CORPUS

The methods to be applied here are quite similar to the ones applied for the identification of lexical units, see above Section 4.1. The first strategy is to export all words of the target language (defined by blanks in the text corpus) and list them alphabetically. The inflected and derived word forms will show up in the list if the stems remain formally constant and have no prefixes. One has to look for systematic variations with regard to form-meaning pairings in the tx line and the ft line. If there are systematic form-meaning variations with respect to some affix-like form or with respect to some potential grammatical words, one can proceed with searching these forms in the text tx tier. For instance, in Hocak, progressive aspect is systematically expressed by means of certain verb-auxiliary analytical constructions. Once the construction is discovered, one may search these constructions systematically on the tx tier.

Another strategy could start with prototypical nouns and verb, since grammatical variation can be expected to occur primarily with words of the open classes. These prototypical nouns and verbs have to be looked for in the translation ft tier. Under the assumption that for instance inflected verbs in the translation ft tier correspond to inflected verbs in the text tx tier, one may start with searches for prototypical verbs and nouns in the ft tier. This strategy can be applied also to grammatical meanings: grammatical meanings such as personal pronouns ('I', 'you', etc.), progressive aspects (e.g. 'V-ing'), and so forth can be searched for in the translation ft tier. The results can be compared with what corresponds to it in the tx line of the object language. The problem with this kind of search is that one will not find complete paradigms in the text corpus. Elicitation of forms with a native speaker is much more promising here as a strategy.

The latter strategy can be exemplified with an example from the Hocak corpus. If one want to find out, if progressive aspect (PROG) is marked grammatically in Hocak, one may search for *V-ing* constructions in the translation ft tier. Of course, the range of hits one gets is much wider than just the clear progressive uses of this form, since the *-ing* form in English is not restricted to progressive aspect. So, one has to select from the results of this search the cases which seem to be good cases of a progressive meaning in English and may compare this with the forms in the text tx tier.

A variation of this kind of search is to choose a specific probably frequent verb and search for its progressive construction in the translation tier. For example, a search for "was looking" receives 6 hits in the Hocak corpus with quite interesting results. First of all, one gets the standard analytic construction in Hocak for the expression of progressive aspect (cf. the text in bold in the second line of example (7) corresponding to the English meaning 'listening to them'); it is one possible verb auxiliary construction for progressive aspect; but one gets also an alternative expressions such as the reduplication (cf. first line in example (7) which corresponds to 'was looking'), and more astonishing, a construction with the habitual marker (glossed HAB) on the verb, cf. (8).

(7) ref TWI026

tx	Hegü	hegü	han'qac	horoğóğoc	hegü
mo	hegü	hegü	hanaqac	horo<ğö>ğoc	hegü
gl	that.way	that.way	all	<RDP:ITER>look.at(SBJ.3SG)	that.way

allomorphs	grammatical meanings/functions	gloss
- <i>kje</i>		
- <i>kjane</i>		
- <i>kjene</i>	future intentional, irrealis, desiderative, obligation	FUT
- <i>kjanqhe</i>		
- <i>kjenehe</i>		

TABLE 3.: Allomorphs of the future tense marker –*kjene* FUT in the Hocak corpus

tx wan'axgu wa'uqkšqna hegu
 mo wa-naaxgu wa'u-'ak-šana hegu
 gl OBJ.3PL-hear(SBJ.3SG) do/be-POS.HOR-DECL that.way
 ft It was looking all around and listening to them.
 dt 21/Sep/2006

(8) ref HOR029

tx Hotoğocnaga hegu šijcrašge nqasura... hoipini
 mo hotoğoc-naga hegu šijc-ra-šge nqasu-ra hoipini
 gl look.at\1E.A-and that.way rear-DEF-also head-DEF spin

 tx haakirinak hegu hakjopahišge
 mo ha<ha>kirinak hegu hakja-ho-hapahi-šge
 gl <1E.A>land.on that.way back-APPL.INESS-go.toward-also

 tx hamianaknagašge, cqaqaja guagaira horoğocnijsgeşunu.
 mo ham<ha>nak-naga-šge cqaq-'eeja guagaira horoğoc-nijsge-şunu
 gl <1E.A>sit.on-and-also outside-there sometimes look.at-VAGUE-HAB
 ft I got on it and spun around on it **glancing** outside once and a while, while I **was looking** over the back end.
 dt 22/Sep/2006

Especially the latter result is due to the polysemy of the English progressive aspect on –*ing* which obviously can be used to express habitual meanings too. The latter type of search equally belongs to the onomasiological approach to grammar.

4.2.2 WITH A MORPHEME-BY-MORPHEME SEGMENTATION AND GLOSSING IN THE CORPUS If grammatical morphemes are already glossed, it is of course easy to find all its allomorphs and all its meanings. Just look for the grammatical gloss on the gloss tier. For instance, if one looks for the Hocak future tense marker glossed FUT one finds at least five different forms (allomorphs) for this marker –*kjene* 'FUT', cf. Table 3.

The results of this search show that there is some allomorphy associated with this morpheme and, in addition, that this morpheme is actually polyfunctional. It marks future tense, but also different kinds of modal meanings such as 'should' and others.

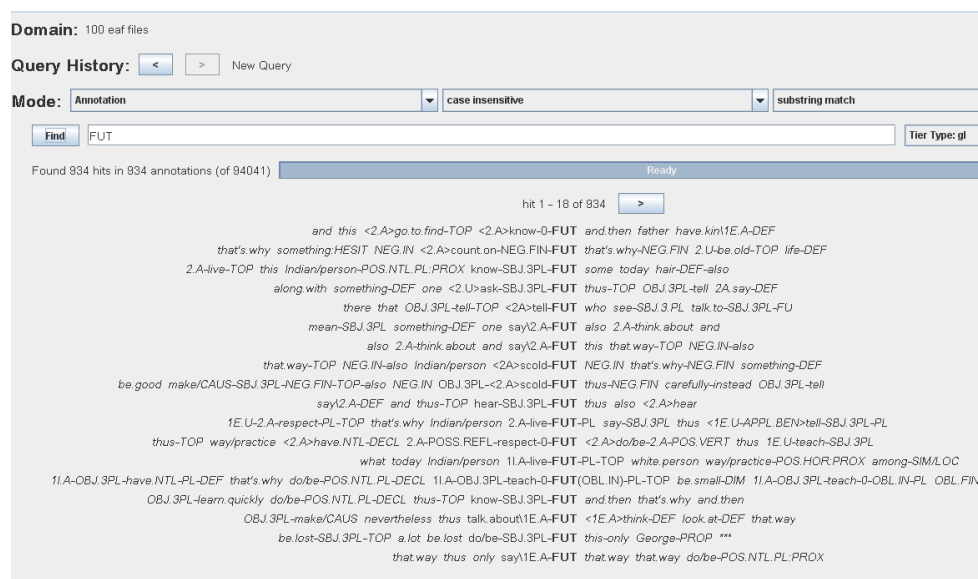


FIGURE 4.: Selection of hits of a search for the gloss FUT (future tense marker) in the gl tier

4.3 WORD INTERNAL MORPHOLOGICAL STRUCTURE – THE COMBINATION OF LEXICAL AND GRAMMATICAL MORPHEMES

The task here is to find out the possible combinations of lexical and grammatical morphemes within the word; in case that clitics are involved the searches have to transcend the word boundary. This can be done in a DOBES corpus with an interlinear glossing simply by searching the grammatical item and analyzing the forms that appear to the left or to the right in the concordances. This search can be executed on the morpheme break mo tier or on the gloss gl tier. This kind of search can be illustrated with an example from Hocak.

Hocak verbs have quite complex chains of suffixes and enclitics which are all formally identified. However, the possible syntagmatic combinations are not known. In order to find this out, all kinds of combinations of grammatical suffixes/ enclitics [stem-X--Y] with all kinds of morphemes in between have to be searched. Because of the morpheme glossing in this corpus, it is easy to find all syntagmatic combinations of the relevant grammatical formatives. It suffices to search for the gloss and then analyze the chains of suffixes/ enclitics. For instance, if one searches for the gloss FUT which is the gloss for the future tense marker *-kjene* 'FUT' and its allomorphs, one gets chains of glosses like the ones given in Figure 4. As can be seen already from this selection, the variety of grammatical formatives that may follow the FUT marker or that precede it is remarkable. In order to arrive at a kind of morphological template for the suffixes/ enclitics in Hocak verbs, various combinations have to be searched for.

Another question would be: how to examine morphophonemic processes with corpus linguistic methods in DOBES corpora. Hocak has heavy morphophonemic processes in the prefixes, i.e. the outcome of the combination of two prefixes often results in a very opaque form. See, for instance, *hıyuroğoc* 'he was looking at me' which appears in example 9. The

underlying form is: *ho-i-Ø-roğoc* with a divided stem *ho-roğoc* 'look at' and the pronominal infixes 1SG.UG(me)-3SG.A(he).

(9) ref HOR068

tx	<i>Hegu</i>	<i>wogitekji</i>	<i>hųųroğoc</i>
mo	hegü	woogitek-xji	ho<i-Ø>roğoc
gl	that.way be.angry-INTS <1E.U-3SG.A>look.at		
tx	<i>wa'uqkşqna,</i>	<i>hegü</i>	<i>'eeja nųųgiwqkji</i>
mo	wa'ü-'ak-şana	hegü	'eeja nųųgiwqk-jî
gl	do/be(SBJ.3SG)-POS.HOR-DECL that.way there run-INTS		
tx	<i>kirikere</i>	<i>haa.</i>	
mo	kiri-kere	haa	
gl	arrive.back.here-go.back.there make/CAUS\1E.A		

ft He was looking at me real mad and I left there running fast.
dt 25/Sep/2006

What happens here is that the underlying pronominal affixes merge with the first part of the stem in quite unpredictable ways. The morphophonemic process can easily be examined by comparison of the gloss in the gl tier with the form in the text tx tier. In order to find out, whether this type of morphophonemic process happens regularly, one may search for the string *hųų* in the text tx tier to get all instances of this process. Unfortunately, this search gives also words that are not inflected in this way such as *hųųc* 'bear' and *hųųk* 'chief'. A solution would be to search across tiers, the form *hųų* itself in the tx line, and the function 1SG.UG-3SG.A in the gl tier, in order to avoid to get all words that begin with *hųų*.

4.4 BEYOND THE WORD LEVEL Structural units beyond the word level of complexity comprise grammatical categories that are expressed analytically, the internal structure of different phrase types such as the NP, the VP, the PP and so on, the constituent structure of the clause and the structure of complex clauses. The descriptive task of a structural grammar is to describe the internal structure and the distribution of these syntactic categories and constructions. For instance, in order to describe the NP as a syntactic category one has to find out the kinds of elements that may be combined in a NP and their possible order(s) within this constituent type. The problem for the corpus linguistic methods is that DOBES corpora do not have a syntactic annotation. Neither parts of speech nor phrases such as NPs etc. are annotated. Therefore, the descriptive linguist is forced to formulate indirect search routines in order to identify the intended phrase type, e.g. a NP. For instance, in Hocak, most NPs end with a determiner on the right edge. Searching for these determiners (e.g. definite and indefinite articles, demonstrative pronouns) could be a way to find NPs in the text corpus. However, many subordinate clauses have a definite article on the right edge too, so they have to be filtered out by a second procedure. Another problem would arise that there are also NPs that have no determiners in this slot. Another possibility would be to search for nouns assuming that they always occur in NPs. This is not possible, since parts of speech are not annotated. In addition, there are NPs without nouns as heads.

So, it can be concluded that the lack of a syntactic annotation hinders the corpus linguistic treatment of DOBES corpora quite a bit. It would be desirable to have a tool for the

automatic or semi-automatic syntactic annotation of DOBES corpora to make them apt for the searches the descriptive linguist needs to analyze the syntax of the target language. This holds for the phrase level, the clause level and the sentence level of syntagmatic complexity.

Another question are grammatical categories that are expressed analytically. In principle, there is the possibility to search for one element of such a construction and then to compare the hits which contain the second element of such a construction with the ones in which this element does not occur. The other possibility is to look for the whole construction, i.e. a combination of two elements that belong to the construction. In the search string one has to specify the two elements by strings of characters and the number of elements that may interrupt this construction by variables (with certain regular expressions). For instance, in Hocak, the modal meaning 'must not' is expressed by means of a combination of the future suffix – *kje* and a separate word *heesgé* following the verb. Both elements occur independently with a specific meaning, but in combination they have the meaning 'must not' which cannot be composed out of the elements. In the Hocak corpus, this construction has already been identified and was glossed accordingly, cf. (10):

(10) ref ALV022

tx *Ciirášge* *hižá paagáxwigi* *hegu* *wažá*
 mo *cii-ra-šge* *hižá paagax-wi-gi* *hegu* *wažá*
 gl house-DEF-also one write(draw)\1E.A-PL-TOP that.way something

tx *hijqhínq* *hižá paagáxwigi* *mąixeté* *cii*
 mo *hijqhí-ra* *hižá paagax-wi-gi* *mąixete* *cii*
 gl different-DEF one write(draw)\1E.A-PL-TOP white.person house

tx *paagáxikjawi* *heesgera* *haqké heesge*
 mo **paagax-i-kje-wi** **heesge-ra** *haąke heesge*
 gl **write(draw)\1E.A-0-OBL.IN-PL OBL.FIN-DEF** NEG.IN that's.why

tx *hawiní* *'eegi* *jaagu waací*
 mo *haa-wi-ní* *'eegi* *jaagu wa-ha-cii*
 gl make/CAUS\1E.A-PL-NEG.FIN and.then what OBJ.3PL-1E.A-live

tx *hanqkwigi* *heesge* *paagáxwigi*
 mo *ha-na-k-wi-gi* *heesge* *paagax-wi-gi*
 gl 1E.A-POS.NTL.PL-PL-TOP that's.why write(draw)\1E.A-PL-TOP

tx *woogitékire*.
 mo *woogitek-ire*
 gl get.angry/mad-SBJ.3PL

ft When we drew a house, **we were supposed to draw the white man's house**, if we drew what we were living in or something different they would get mad.
 dt 23/Jul/2007

However, if this is not the case, one has to formulate a complex search or to formulate subsequent searches in order to identify constructions like these.

5 FUNCTIONAL GRAMMAR – THE ONOMASIOLOGICAL APPROACH The goal of the onomasiological approach to grammar is to describe in a systematic way the linguistic means by which in the target language a certain general function can be expressed. This approach to grammar requires searching for semantic functions/ concepts and operations that are considered to be necessary more or less for each language. They are therefore considered universal. It is self-evident that there cannot be a complete list of such functions. Table 4 presents a summary of the most important (probably universal) tasks a language has to be able to fulfill.

Table 4 contains general functions such as reference, possession, and spatial orientation which are fulfilled by various formal means cross-linguistically, but also intra-linguistically. The latter means that even in an individual language there are different means by which these functions are expressed formally. However, these general functions are too abstract to be searchable in the corpus, since these general functions are not or only indirectly annotated in the glossing tier, or are not or only indirectly coded in the translation language English in the translation line. In order to be able to search these general lexical or grammatical meanings/ functions, one has to find lexical items or construction in the translation language – in the Hocak corpus, it is English – that express the intended meaning/ function or are at least cues for constructions that express the intended function. In the right hand column in Table 4, lexical expressions and grammatical construction in English are enumerated that may serve as search cues for certain semantic concepts and functions. How one can search certain linguistic phenomena that are cues for certain semantic concepts and functions will be illustrated in the subsequent sections Section 5.1 and Section 5.2. The possibility to search for semantic concepts and function utilizing the information in the translation tier is the major advantage a parallel corpus like the ones of the DOBES projects has over traditional monolingual corpora.

5.1 TIME IN HOCÁK In order to find out how tense is expressed in Hocak, one has different possibilities to search in the English translation language. One may search for tensed verbs in English such as the past tense form of verbs or the future category expressed by a verb-auxiliary construction. In addition, one may search for temporal adverbials that indicate the time relation of the English clause assuming that in the corresponding Hocak clause, there will be a similar time adverbial.

For instance, if one searches the standard future construction in English, the *will* plus verb construction it suffices to search the string “*will*” in the translation tier. One gets 314 hits of “*will*” in the Hocak corpus. The corresponding Hocak clauses can then be looked at to see if there is a corresponding grammatical marker indicating future tense. In almost all cases, one can find one of the above mentioned allomorphs of the future morpheme *-kje*. An additional possibility is to conduct a combined search across tiers like “*will*” in the *ft* tier and NOT *FUT* in the *gl* tier, in order to find out, if the hits in English correspond always to the future marker (*FUT*) in Hocak. Interestingly, there are cases when we find a potential marker *-nqq* in Hocak corresponding to the *will*-construction in English; cf. (11).

domain	basic functions	representative concepts and operations	lexical expressions and grammatical constructions in English as cues for certain semantic concepts
apprehension & nomination	an entity is grasped by categorizing and individuating it; it is named by a label or a descriptive expression	categorization, types of concepts, empathy	NUM-N N-SG/PL
concept modification	a concept is enriched, or an object is identified	attribution, apposition, relativization	ADJ-N, PTC-N, N-REL constructions, PRO/PROP-appositions
reference	a representation is related to and delimited within the universe of discourse	determination, deixis, reference tracking	DEM-N, DEF-N, INDEF-N, SAP-PRO, Non-SAP 3 rd person pronouns
possession	the relation of an entity to another one is established or inheres in one of them	possession in reference, possessive predication, external possessors	genitive attribute (<i>-s/ of</i>) verbs of possession (<i>have, belong</i>) dative of possession (e.g. <i>carry sth. for so.</i>)
spatial orientation	an entity is localized in space statically or dynamically	reference points, local relations, spatial and gestalt properties of objects	spatial adverbs N ₁ PREP-N ₂ constructions with N ₁ as the Figure (object localized) and N ₂ as the Ground (region/ frame of reference)
quantification	the extent of the involvement of a set of entities in a predication is delimited	quantification in reference and in predication; counting, ordering	N-SG/PL NUM/QUANT-N numeral adverbials cardinal and ordinal numbers
predication	information is attributed to a referent	existence, situation, characterization	X is Y X is a Y/ Xs are Ys there is a X X is PREP NP X is POSS.PRO Y and many more constructions of location and possession
participation	a situation is articulated into an immaterial center and a set of participants and circumstants related to it and to each other	control & affectedness, central vs. peripheral roles, alignment of fundamental relations	Semantic types of verbs as cues for different event types, verb classes

temporal orientation	a situation is designed with respect to its internal temporal structure and limits and temporally related to another situation	situation types, aspectuality, temporal relations	perfective vs. imperfective aspect (only indirectly in English, but see the progressive aspect) verb classes (stative vs. dynamic; cf. test with progressive) tense (past (V-ed) , present perfect (have/has V-ed) , present (0/ -s) , future (will V)), temporal adverbials (<i>yesterday, tomorrow, last month, next Monday</i> , etc.)
illocution, modality, evidentiality	a proposition is rendered relative to speaker, hearer and reality	speech acts, obligation, volition, possibility, toning, evidentiality	question words, question marks, verb first position as indicator for imperative clause, exclamation mark modals such as <i>want, like, must, have to, should, may, can</i> , etc.
contrast	a concept or proposition is assessed qualitatively by comparison with similar ones	negation, comparison, gradation, intensification	referent negation (NEG-N), predicate negation (NEG V) comparison (positive, comparative, superlative of the ADJ) DIM/AUG
nexion	a situation is expanded into a complex one, or several situations are linked together	speech reproduction, complementation, interpropositional relations	coordinating and subordinating conjunctions (such as <i>and, but, while, because</i> etc.) as cues for interpropositional relations, coordinate clauses and subordinate clauses
communicative dynamism	a proposition is articulated in foreground and background	discourse structure, functional sentence perspective (topicalization, focusing, emphasis)	there is a X that ... It is X that ...

TABLE 4.: Functional Domains and the formal cues in English (based on Lehmann 2004).

(11) ref WIC008

tx *Taawusiregi* *wiicqt'jiranqanq*,
 mo *taawus-ire-gi* *wiicqt'i-ire-naa-na*
 gl be.dried-SBJ.3PL-TOP be.noticeable-SBJ.3PL-**pot**-DECL
 tx *haašak* *hiiranqanq*.
 mo *haašak* *hii-ire-naa-na*
 gl be.crusty(OBJ.3SG) make/CAUS-SBJ.3PL-**pot**-DECL
 ft When they are dry, they **will** become tough, hard like a shell.
 dt 07/Jun/2005

The use of the potential marker in Hocak in this context in (11) makes perfectly sense, since the context is a conditional clause expressing a strong probability in the irrealis and not a future event. The English will construction is obviously not sensitive for this categorial distinction.

Similar search procedures can be performed for the other tense categories. For instance, to search for *V-ed*, one may search for the string *ed\b* in the regular expression mode (in Elan) in the ft tier and get about 3000 hits. Looking through the tx and mo tier of the hits one can see that there is no past tense marker in Hocak.

5.2 COMPARISON IN HOCÁK Another illustrative example for a fruitful application of the onomasiological search strategy would be comparison. Hocak has no grammaticalized comparative and superlative. But, how can one find out, how these concepts that are tightly bound to adjectives are expressed in Hocak. Again, one can look into the ft tier in order to find comparative or superlative forms which can be compared with the corresponding expressions in Hocak. The cues for comparative to look for in English are the ending *-er* and the marker of the standard of comparison *than*; e.g., *fast-er than his horse*. The *-er* alone is not a good cue for the comparative, because one gets all words that end in *-er* (n=1820 among them *father*, *after* etc.) and these are mostly not comparatives. Hence, *than* would be the better cue. Not surprisingly, there are only 7 instances in the whole corpus. It seems this is a category that is strongly avoided. The strategies to express comparison in Hocak are lexical as can be seen for instance in (12).

(12) tx *hųćrá šųćkj'akra hirakísanjk hųćrá 'ee wamašć'qanq*
 mo *hųć-rá šųćkj'ak-ra hirakísanjk hųć-rá 'ee wamašć'a-na*
 gl bear-DEF wolf-DEF **compared** bear-DEF **he.EMPH** strong-DECL
 ft 'The bear is stronger **than** the wolf' (lit. 'The bear **compared with the wolf**, the bear, **HE** is strong.')

In tx *hųćrá šųćkj'akra hirakísanjk hųćrá 'ee wamašć'qanq*, the asymmetry between the bear and the wolf with regard to the dimension strength is expressed by means of a clause that indicates that there is a comparison and by means of a subsequent clause that focuses the entity that has more of the dimension. This is not a grammaticalized construction, but a possibility speakers invent ad hoc in order to express the concept. There are many other different possibilities to express comparative.

A similar situation can be found in Hocak with regard to the superlative. One may choose as cue in English the ending *...est/b* for a search in the ft tier. Of course, there are words in English that end in *...est* like *west* or *rest* that are not relevant. The result would be that a) the superlative occurs very rarely, and b) that it is not a grammaticalized construction. One possibility to express a concept like the superlative in Hocak would be lexically such as in 13.

- (13) tx wqak hakjnúpra waihakra 'ee
 mo wák ha-kjnúp-ra wa-0-hihak-ra 'ee
 gl male coll-sibling-DEF 3PL.OBJ-3SG.A-on.top-DEF he.EMPH
 tx hereénq, Adam Little Bear Jr., raašrá 'Aahú Ru'qgá
 mo hereé-nq, Adam Little Bear Jr., raaš-rá 'Aahú Ru'q-gá
 gl be-DECL Adam Little Bear Jr., name-DEF Wing Raising-pn
 tx higaíreenq
 mo higa-íree-nq
 gl call-they-DECL
 ft 'My younger brother, Adam Little Bear Jr., whose name was Raising Wings, is **the youngest** in the family' (lit. translation: 'Of the male siblings, HE (my younger brother) was **on top of them**, Adam Little Bear Jr. he was called Raising Wings.'
 (cf. White Eagle 1988:v))

The meaning of the superlative is that an entity has - with regard to a certain class of the same or comparable entities - the highest value on a certain dimension. In tx wqak hakjnúpra waihakra 'ee hereénq, it is expressed that Adam Little Bear is the youngest in the family, i.e. with respect to the dimension youngness, he is the one who has the highest value on this scale. What is astonishing here is that it is the property youngness that is the reference scale. This scale is invoked in the previous context of this text segment, where the speaker speaks of the younger brother using the special term for this in Hocak.

In the subsequent section, we want to review and evaluate the searching and concordancing functionality of Elan and Poio in order to arrive at list of requirements a software has to offer in order to support semasiological and onomasiological analysis of corpus data.

6 SOFTWARE-BASED SEARCH AND ANALYSIS OF CORPUS DATA In order to be able to evaluate software tools for linguistic analysis we use the same approach taken in Nordhoff (2008) and define a set of values and maxims. Values in this sense define certain criteria to judge the quality of a software from a user's perspective. As Nordhoff states "some of these values can be conflicting", which means that different users prefer different solutions. Still, we consider most of the following values relevant for our purpose of comparing the two packages Elan and Poio (and others not mentioned here, of course) for search and analysis tasks in linguistics. The values are presented as a list of numbered maxims that will later be used to judge whether or not a software tool is better regarding a given value. So the maxims presuppose a certain point of view regarding the values. The reader might not follow all of our assumptions, but all the maxims are derived from the practical application of the ideas to develop a descriptive grammar based on corpus analysis presented in this paper. Our maxims are:

1. Search results should be presented as interlinear text.
2. The user should be able to find the source utterance in its context in the original file from the search result.
3. The user should be able to search on all existing tiers.
4. Relationships among search terms:
 - a) It should be possible to define relationships among search terms on one tier.
 - b) It should be possible to define relationships among search terms on different tiers.
5. The user should be able to search within search results.
6. Search should be possible in a set of files, not only in one file. The more file formats supported, the better.
7. The user should be able to search for substrings in annotations and use regular expressions.
8. It is better when the user is confronted with fewer dialogs and windows during one search tasks.²
9. It should be possible to export searches and search results, in order to save and archive them for later reference.

The following two sections first give an overview of both software packages, specifically of the search and analysis functionality implemented in the tools. Each section will conclude with a summary that contains a rating of the tool according to the maxims we outlined above.

7 REVIEW OF ELAN SEARCH AND ANALYSIS FUNCTIONALITY One of the widespread tools in language documentation nowadays is Elan, a transcription and annotation software developed by the Max-Planck-Institute in Nijmegen, Netherlands.³ As Elan was developed specifically for transcription and annotation of audio and video files in the beginning, search and analysis functionality was only added later to the software. This led to the current situation, where this functionality is distributed among several menu options within the user interface. Regarding the search types and approaches to corpus linguistic methods for descriptive purposes presented in the first part of this paper, Elan offers the following three features that fulfill some of the requirements the descriptive linguist is striving for:

- Export multiple files as: list of words/annotations
- Search multiple eaf (files)
- Structured search multiple eaf (files)
 - Substring search

² This is only an approximation of the quality of the graphical user interface design; a full review of the usability of each tool is left for future research.

³ <http://www.lat-mpi.eu/tools/elan/>

Word form	frequency	Word form	frequency
xunu	1	xunu-ig-naagre	1
xunu-ig	1	xunu-ig-ra	1
xunu-ig-iža	1	xunu-ik	6
xunu-ig-na	1	xunu-ik-ra	2

TABLE 5.: Extract from the word list of the entire Hocak corpus

- Single layer search
- Multiple layer search

The “Structured search multiple eaf” feature is by far the most complex and elaborated. It is in itself divided in three dialogs which are listed above. All of the features work on what is called a “domain” in Elan. A domain in this sense is a batch of files. The user can create a domain, give it a name, and then add Elan files to that domain. Later, within each search and analysis feature, users first select a pre-defined domain and then carry out their search and analysis on that domain. The user is able to do several types of analysis through all of the listed features. The following sections will each describe one of the features, and relate their functionality to the search types that were described in Section 3.

7.1 EXPORT MULTIPLE FILES By the “Export multiple files” feature, the user can export all words or annotations of a batch of files to a simple text file. Figure 5 shows the export dialog for word lists. The user selects the tiers from which to export and defines a token delimiter. The default option is to use the pre-defined delimiters, which tokenize by punctuation. Another option is the frequency count (“Count occurrences” option in the dialog), which is added tab-separated to each entry in the exported text file. If the user chooses the frequency count each word in the export file is then accompanied with a number how often this word occurs in all of the domain files.

Figure 5 shows an extract of a sample export file of the Hocak corpus. This method allows to some degree the identification of allomorphs of the word/lexeme as mentioned in Section 4.1. The extract shows a list of words with a common prefix *xunu* ‘small, little’. Prefixed forms appear elsewhere in the list and have to be searched separately. Alternate stem vowels (long vs. short) will also lead to a separation of list entries with common lexical roots. The user is not able to trace back the entries to the corpus. Once the export is done, the user cannot simply jump to the position within the files where a list entry occurs. The export of the full interlinear context of words/annotations is not possible with Elan.

7.2 SEARCH MULTIPLE EAF The second feature to be discussed here is the “search multiple eaf” option. This search functionality allows the user to search in the domain’s eaf files for search terms. Search terms can be regular expressions or raw strings. The only other option in the search dialog is to switch on or off case sensitivity. Figure 6 shows the search result dialog for the search term *xunu* ‘small, little’. In this case, the user is able to jump directly to the occurrence of the search term by double-clicking on any entry in the result list. In addition, the user is presented with the annotations before and after the search hit. This

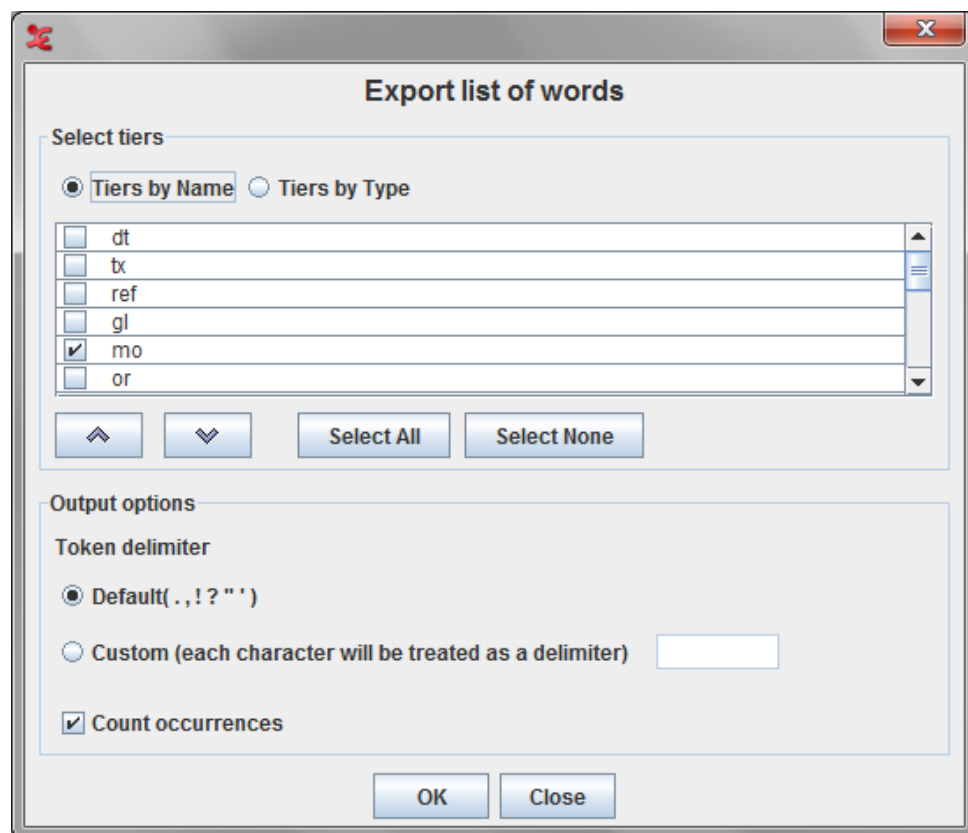


FIGURE 5.: Word list export in Elan

Nr	File	Tier	Before	Annotation	After	Begin Time	End Time	Duration
1	05IH_Richard...	mo	hoc[c]-jk	xunu-jk	ha-ni-he-regi	00:00:00.810	00:00:04.849	00:00:04.039
2	05IH_Richard...	mo	cinak	xunu-jk	nijisge	00:02:37.386	00:02:41.281	00:00:03.895
3	05IH_Richard...	mo	nijkjak	xunu-nijisge	ha-ni-he-nak-ga	00:04:03.374	00:04:12.418	00:00:09.044
4	05IH_Richard...	mo	'eeja	xunu-jk	ha-ni-he-regi-giži	00:04:12.418	00:04:17.574	00:00:05.156
5	della_CORR	or	žeeğu haagiži ...	haha, 'eeekjana 'eegi	haha 'eesge ...	00:10:55.982	00:11:02.742	00:00:06.760
6	fox_war	cm	lit. However ...	majisic njne = ...		00:06:05.000	00:06:10.000	00:00:05.000
7	JF04	mo	cinak	xunu-jg	here	00:01:01.175	00:01:15.695	00:00:14.520
8	JF06	mo	kook	xunu-jg-naagre	heesge-ra	00:01:08.283	00:01:20.665	00:00:12.382
9	JF06	mo	kook	xunu-jg-na	žeeğu	00:01:08.283	00:01:20.665	00:00:12.382
10	JF06	mo	kook	xunu-jg-ra	kerepana	00:01:32.158	00:01:43.160	00:00:11.002
11	JF06	mo	jaagu	hi-xunu	wa'u	00:04:48.266	00:04:55.920	00:00:07.654
12	jones	mo	wa-haa-ra	xunu??-jk??-ga??	Heena-ga	00:01:38.989	00:01:46.198	00:00:07.209
13	ken_pauline_IH...	mo	wooxja_hii-ire-ga	hi-xunu-xj-jk	hi-xunu-xj-jk	00:08:04.652	00:08:22.831	00:00:18.279
14	ken_pauline_IH...	mo	hi-xunu-xj-jk	hi-xunu-xj-jk	ho<hi>kife-ire...	00:08:04.652	00:08:22.831	00:00:18.279
15	ken_pauline_IH...	mo	'eegi	hi-xunu-xj-jk-šana	'eegi	00:08:52.052	00:09:00.345	00:00:08.293
16	ken_pauline_IH...	mo	'eegi	xunu-xj-jk	hegu	00:09:17.880	00:09:29.616	00:00:11.736
17	ken_pauline_IH...	mo	hegu	xunu-xj-jk	hegu	00:09:29.616	00:09:45.292	00:00:15.676
18	ken_pauline_IH...	mo	anaga	xunu-xj-jk	hegu	00:11:06.537	00:11:13.939	00:00:07.402
19	ken_pauline_IH...	mo	jaagu	xunu-jk	wa'u-ha-jee	00:12:23.223	00:12:27.616	00:00:04.393
20	ken_pauline_IH...	mo	paaši-šunu	xunu-xj-jk	hegu	00:13:25.503	00:13:34.062	00:00:08.559
21	ken_pauline_IH...	mo	nige	xunu-jk	ni-šge	00:27:40.082	00:27:48.170	00:00:08.088
22	ken_pauline_IH...	mo	paaxge	xunu-g-iža	žige	00:37:16.783	00:37:20.523	00:00:03.740
23	ken_pauline_IH...	mo	hejaga	xunu-šeeep-jk	hija	00:38:49.190	00:39:06.298	00:00:17.108

FIGURE 6.: Search multiple eaf in Elan

search is applied to all tiers in all files. It is not possible to restrict the tiers to search. The user can export the result list in a tab-separated format.

The search results are not displayed in full interlinear context, only a user-defined context on the same tier before and after the search term is shown.

7.3 STRUCTURED SEARCH MULTIPLE EAF Regarding the search typology presented in Section 3 above, the “structured search” is the most comprehensive facility for search and analysis in Elan. The “structured search” is in itself divided in three dialogs, hence search facilities: a simple substring search over all domain files, the single layer search and the multiple layer search. As only the latter two go beyond the “search multiple eaf” option already presented in the previous section, we will be only concerned with the search in layers here.

Figure 7 shows both search dialogs. The left dialog presents a search for *xunu* in all tiers of the type mo. The “multiple layers search” is shown in the right dialog, with sample search for *xunu* in all tiers of type mo together with an occurrence of *little* in tiers of type ft. In this example, the constraint between the tiers is set to “No overlap”. The green boxes between search term boxes in the dialog allow the user to set different constraints within and between tiers. Constraints between tiers are concerned with different types of overlap, for example “Left overlap” or “Within”. Constraints within a tier set the a maximum, minimum or exact distance of two search terms within the given tier, in terms of annotation counts or milliseconds within the media file.

Both dialogs present the results in the lower parts of the dialog as concordances. The user can define the context size of the concordance. An alternative view is given by the

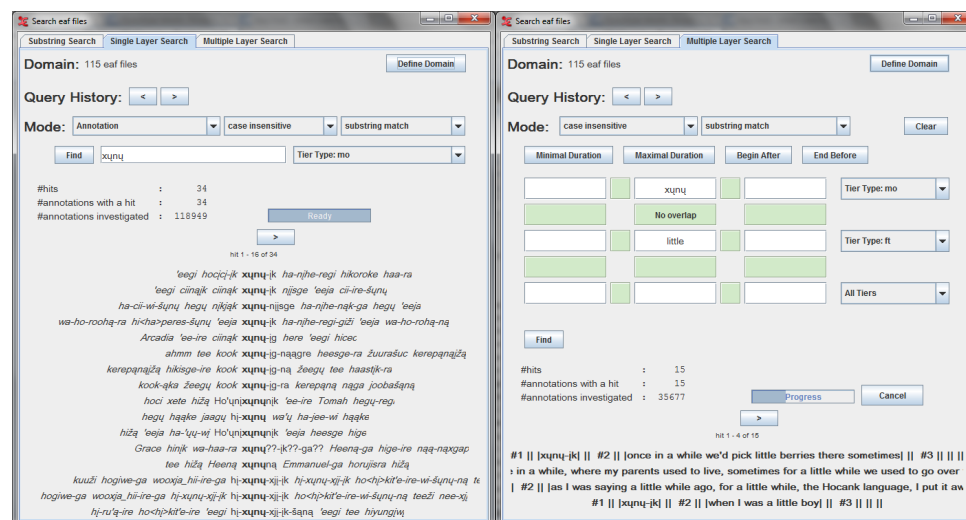


FIGURE 7.: Structured search in Elan

“frequency view”, in which all search hits are listed in a table with a count and a percentage. In both views the user can jump directly to the annotation in the eaf file by double-clicking on the element in the list view. The search results may also be exported to a tab-separated file with all timing and context information.

7.4 SUMMARY Regarding the maxims in Section 6, the software Elan is rated as follows:

1. **Search results should be presented as interlinear text:** It is not possible to view search results as interlinear text in Elan. Elan does not even display the annotation for all the tiers at the location where search strings occurs. Only the context within the tiers that was searched in is displayed. Elan does ignore this value completely.
2. **The user should be able to find the source utterance in its context in the original file from the search result.** The user is presented with a configurable context within the list of search results in the simple and complex searches. The user can click on each element in the list and Elan will open another window with the file of the search hit. The user then has to scroll left and right to see the context of the annotation and utterance within the file. We see problems in usability here, but in general Elan implements this feature.
3. **The user should be able to search on all existing tiers:** Elan does allow searches on each tier without restrictions.
4. **Relationships among search terms:**
 - a) **It should be possible to define relationships among search terms on one tier;**
 - b) **It should be possible to define relationships among search terms on different tiers:**

Elan allows both types of connections between search terms. For some of our users in the project the connection names like “overlap”, etc. were not very clear, so we think that usability could be improved. On the other hand the detailed connection types allow every thinkable search combination for the expert user.

5. **The user should be able to search within search results:** Elan does not allow any further searches on the search result.
6. **Search should be possible in a set of files, not only in one file. The more file formats supported, the better.**
Elan support searches in multiple files through its “domain” concept. But only one file format is supported directly, Elan’s EAF format. The user may search in other file types by importing the files first and save them as EAF files.
7. **The user should be able to search for substrings in annotations and use regular expressions.** Elan does fully support regular expressions.
8. **It is better when the user is confronted with fewer dialogs and windows during one search tasks.**⁴ Elan’s search functionality is scattered among several menu options. So, the user has to open several menus and dialogs to try every possible search type. In the case of the export of word lists the user even has to open another application to apply the search. In our view Elan fails regarding this maxim.
9. **It should be possible to export searches and search results, in order to save and archive them for later reference.** It is possible to export the search results in Elan, but not the searches.

In summary, Elan implements features for 6 of the 9 maxims. Maxims 1, 5 and 8 are violated completely, maxims 4, 6 and 9 only partly or only when we additionally take usability into account (in case of maxim 4). Elan’s strength definitely is the structured search that allows advanced search strategies if the user learns about the connection types. Definitely missing is the interlinear view of search results, we also found major restrictions in getting context information from search results. Although the latter is possible, it is quite difficult for the user to get a quick overview of the search result and its context on all tiers at one glance.

8 THE POIO ANALYZER FOR DESCRIPTIVE LINGUISTS The development of the Poio Analyzer was started because we felt that Elan was not fulfilling all our needs. The software was planned as a tool for descriptive linguists working with data from language documentation projects from the beginning. Once the search typology emerged during the analysis of the Hocak data, it became clear that Elan is not the right tool for a grammar writer to carry out the analysis. The development of the Poio Analyzer did not start from scratch though, as there was already a library for Toolbox and Elan file access (PyAnnotation⁵) and

⁴ This is only an approximation of the quality of the graphical user interface design; a full review is left for future research.

⁵ <http://www.cidles.eu/ttl/poio-pyannotation>

a morpho-syntactic annotation editor (Poio ILE⁶) available. The first provided a common API to the data files with access to the morpho-syntactic annotation, while the latter already contained a view to display utterances in an interlinear style. Both software packages were developed and used within the DOBES project “Minderico –An endangered language in Portugal”.⁷ All software packages including Poio Analyzer are available for download on the Poio website⁸ and are licensed under the GNU General Public License v3.0. Besides the source code there are installation packages for Windows available.

The following section will describe the user interface of Poio Analyzer, with a focus on the requirements that are missing in Elan and that Poio Analyzer implements. We will then give an outlook on future features that are currently work-in-progress before summarizing Poio’s features and rating the software.

8.1 THE USER INTERFACE OF POIO ANALYZER Figure 8 shows a screenshot of the Poio Analyzer. The main sections of the GUI are marked by numbered boxes. The example data in this screenshot show a search for the regular expression “`^ra$`” on the morpheme tier, the gloss “DEF” on the gloss tier and the word “`\bthe\b`” within the translation. The functionality of each part of the GUI surrounded by boxes is as follows:

1. Search tabs allow successive searches. Each “New Search...” tab filters on previous searches. Tabs allow changing intermediate searches by switching between the open tabs. Results are displayed for the current tab.
2. Input fields for search strings to search in each tier. Allows full-flavored regular expression through the updated Python regex module. E.g. character classes like `\w` or `\b` fully support Unicode; there is support for Unicode properties, blocks and scripts, etc.
3. Search options:
 - a) AND: matches, when all search strings of all tiers match.
 - b) OR: matches, when one of the search strings matches.
 - c) NOT: inverts the search result.
 - d) “contained matches”: this value alters the matches in the word, morpheme and gloss tier. If set, the search strings must match within the same word. Otherwise the search strings can match in the whole utterance. For example: if you search for “ra” in the morpheme tier, and for “DEF” in the gloss tier, the normal search will display all matches, where “ra” occurs in a morpheme plus all matches where “DEF” occurs in the gloss tier within one utterance. If “contained matches” is set the result will only contain matches where a word has “ra” in the morpheme tier and “DEF” in the gloss tier.
4. Buttons for searches. Searches may be saved and restored; this will save all the search strings of all tabs.

⁶ <http://www.cidles.eu/ltll/poio-ile>

⁷ <http://minderico.caorg.pt>

⁸ <http://www.cidles.eu/ltll/poio>



5. Here you can add or delete files from the corpus. The result view will be updated whenever the file list is changed.
6. The result view. Matches are displayed in green color within utterances and translations. If there is a match on the word, morpheme or gloss tier the whole word and all its morphemes and glosses will be highlighted in green color. This allows the user to find the matches more easily. All utterances are always displayed with full interlinear annotations.

Electronic Grammaticography

formatives. The biggest advantage of Poio Analyzer here is the view of results as interlinear version.

8.2 WORK IN PROGRESS One of the features that were mentioned in Section 8 was the possibility to archive searches and result sets. This feature is currently not available in Poio Analyzer, but we plan to soon release an update that at least allows saving and restoring searches. Other functionality that we are currently working on is:

1. The possibility to search for set of words, morphemes, glosses, etc. This feature seems critical as it allows for example to search for word classes (e.g. part-of speech) which are not directly derivable from the annotations. In language documentation project there are often preliminary, sorted word lists available, that can be used for this kind of search. In the case of Hocak a full dictionary is available, which we will use to search for word classes.
2. On top of this we want to make it possible to add part of speech annotation to the data. In the case of Hocak there is part of speech information available in the dictionary. We are currently investigating the quality of part of speech tagging from this dictionary. Part of speech information would allow search for argument structures as described in Section 4.4.
3. Add statistical evaluation of search result sets. This will be a simple count and percentage value for each tier first. Later, more advanced statistical methods may be added, depending on the needs of the users. The statistical view will also give access to list of words, morphemes and glosses in the result set. Those lists may then be exported and/or used for the “set search” described in 1.

In addition to this we are always collecting ideas for improvement from our users. Depending on the complexity and necessity we will implement anything that seems useful to the descriptive linguist.

8.3 SUMMARY Regarding the maxims in Section 6, the software Poio is rated as follows:

1. **Search results should be presented as interlinear text.** Poio does display all search results as interlinear text.
2. **The user should be able to find the source utterance in its context in the original file from the search result.** Poio does only view the utterance of the search hit. The context of the hit within the utterance is displayed completed, but utterances before and after the search hit are not presented to the user.
3. **The user should be able to search on all existing tiers.** Poio allows searches over all tiers that are relevant for interlinear texts. Additional tiers (that are not part of the hierarchy of a given utterance tier) are not supported yet.
4. **Relationships among search terms:**
 - a) **It should be possible to define relationships among search terms on one tier.**

- b) **It should be possible to define relationships among search terms on different tiers.** Poio does support both types of relationships, but not as elaborated as in Elan. For example it is not possible to restrict the search on a given tier with information about context size ("[search term 1] with a maximum distance from [search term 2]").
5. **The user should be able to search within search results.** This is fully supported by Poio's successive search functionality.
 6. **Search should be possible in a set of files, not only in one file. The more file formats supported, the better.** Poio supports Elan EAF, Toolbox TXT and Kura XML files. Files with different formats may be opened in parallel. The search is carried out over all open files.
 7. **The user should be able to search for substrings in annotations and use regular expressions.** Regular expressions are fully supported by Poio.
 8. **It is better when the user is confronted with fewer dialogs and windows during one search tasks.**⁹
Poio follows this maxim, it only consists of one dialog.
 9. **It should be possible to export searches and search results, in order to save and archive them for later reference.** Poio does currently not support this feature.

In summary Poio implements features for at least 7 of the 9 maxims. Maxims 2 and 9 are not supported, implementation of features is only planned for future releases. In addition, maxim 3 and 4 are partly violated, as Poio does not support searches on all tiers if they are not part of the interlinear text convention. Combinations of search terms have certain restrictions, especially for searches on a single tier. If we take all the maxims into account Poio supports more of the ideas expressed by the values than Elan. It definitely has an advantage regarding usability, as it was developed as a special purpose tool for search and analysis. Elan, as a general purpose tool, has a lot of other features, like the possible to edit the files during analysis. The possibility to combine search terms in advanced ways is currently a standout feature of Elan.

REFERENCES

- Hartmann, Iren, Johannes Helmbrecht, Christian Lehmann & Christian Marschke. 2009ff. *Documentation of Hoocqk*. Nijmegen: Max Planck Institute for Psycholinguistics, DoBeS Archive. http://corpus1.mpi.nl/qfs1/media-archive/dobes_data/Hocank/Corpusstructure/1.imdi.
- Lehmann, Christian & Elena Maslova. 2004. Grammaticography. In Christian Lehmann, Geert Booij, Joachim Mugdan & Stavros Skopeteas (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung*, vol. 17.2, 1857–1882. Berlin & New York: W. de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft).

⁹ This is only an approximation of the quality of the graphical user interface design; a full review is left for future research.

Nordhoff, Sebastian. 2008. Reference Grammars for Typology - Challenges and Solutions. *Language Documentation & Conservation* 2. 296–324.

White Eagle, Josephine. 1988. *A Lexical Study of Winnebago*, vol. 26 Lexicon Project Working Papers. Cambridge, MA: Center for Cognitive Science, MIT.